A3Q4Q5

Tianyu Zhang and Feifei Fu

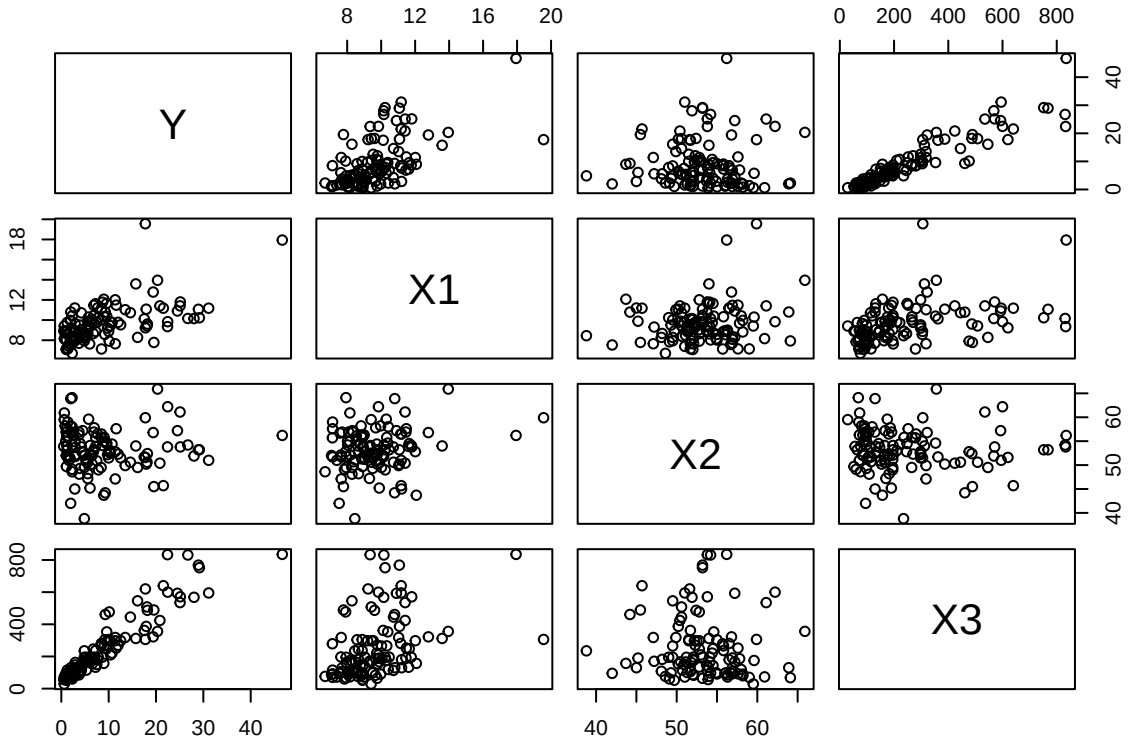
2022/11/9

Import the data from csv file:

```
infection_fulldata <- read.csv("Infections.csv")
names(infection_fulldata) = c("Y","X1","X2","X3")
infect_data = infection_fulldata[,1:4]
```

(a)

```
## Create the scatter plot matrix
pairs(infect_data)
```



```
## Create the correlation matrix
cor(infect_data)
```

```
##            Y        X1          X2          X3
## Y  1.0000000 0.5878580  0.01109340  0.92979478
## X1 0.5878580 1.0000000  0.18891397  0.40926525
## X2 0.0110934 0.1889140  1.00000000 -0.05882316
## X3 0.9297948 0.4092652 -0.05882316  1.00000000
```

```
## Interpret:
```

By the scatter plox matrix: 1. We can see that all the points in scatter plox for Y and X1 are having sort of line patten, so there may have moderate positive linear association between Y and X1; 2. We can see that all the points in scatter plox for Y and X2 are having no patten and is actually a mess, therefore Y and X2 may not having any association; 3. We can see that all the points in scatter plox for Y and X3 are having a positive line patten, so there may have a strong positive linear association between Y and X3;

```
##############################################
```

By the correlation matrix: 1. We can see that from the correlation matrix, corr(Y,X1) = 0.5878580, so Y and X1 are having a moderate positive linear association; 2. We can see that from the correlation matrix, corr(Y,X2) = 0.01109340, it is close to 0, so Y and X2 are having a weak association; 3. We can see that from the correlation matrix, corr(Y,X3) = 0.92979478, so Y and X3 are having a strong positive linear association;

```
##############################################
```

And we don't need to concern about multi-collinearity since the association between X and another X are weak: corr(X1,X2) = 0.1889140, corr(X2,X3) = -0.05882316, corr(X1,X3) = 0.4092652

(b)

```
fit = lm(Y~X1+X2+X3, data = infect_data)
summary(fit)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3, data = infect_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4581 -1.0379 -0.0222  1.1435  7.8990
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11.94805    2.89545  -4.126 7.23e-05 ***
## X1            1.08827    0.13841   7.863 2.90e-12 ***
## X2            0.02568    0.05420   0.474    0.637
## X3            0.03646    0.00135  27.018  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.484 on 109 degrees of freedom
## Multiple R-squared:  0.9163, Adjusted R-squared:  0.914
## F-statistic: 397.9 on 3 and 109 DF,  p-value: < 2.2e-16
```

Estimated regression function: $\hat{Y} = -11.94805 + 1.08827X_1 + 0.02568X_2 + 0.03646X_3$ Interpret $\hat{\beta}_2$: Holding other variables(X1,X3) uncahnge, When age increase 1 year, the average infection will increase 0.02568%.

(c) Step 1: H0: $\beta 1 = \beta 2 = \beta 3 = 0$ Alrernative: Ha: at least one of the $\beta 1, \beta 2, \beta 3$ is not 0 Decision rule :
1. Reject H0 if the test statistic > critical value
2. Reject H0 if the P-value < α

By the summary table in (b), we can see that the P-value is 2.2e-16
Conclusion: Since P-value = 2.2e-16 < $\alpha = 0.05$, we reject H0:$\beta 1 = \beta 2 = \beta 3 = 0$ so at least one of the $\beta 1, \beta 2, \beta 3$ is not 0, there is a linear association between X1,X2, X3 therfore there is a regression relation.

(d) By the summary table in (b), we can see that $R^2 = 0.9163$ and $R^2_{adjusted} = 0.914$ R^2 means 91.63% variation in Y(infection) can be explained by the model(i.e. its linear relationship with X1,X2,X3) And $R^2_{adjusted}$ is a modified measure that accounts for the number of variables in the model. It does not have a actual meaning of interpretation.

(e)

```
set = data.frame(X1 = 10, X2 = 45, X3 = 150)
predict(fit, set, level = 0.9, interval = "predict")
```

```
##        fit      lwr      upr
## 1 5.559544 1.336972 9.782117
```

Interpretation: When the average length of stay of all patients in hospital is 10, the average age of patients is 45 and the average number of beds in hospital during study periods is 150, We have 90% confidence for an infection rate in the hospital is between 1.336972% and 9.782117%.

Q5

(a)

```
infection <- read.csv("Infections.csv")
dataq5 <- infection[,c(1,4,6)]
glimpse(dataq5)
```

```
## Rows: 113
## Columns: 3
## $ Infections <dbl> 8.487, 0.816, 2.214, 2.968, 7.638, 7.497, 6.946, 21.546, 5.~
## $ Beds       <int> 279, 80, 107, 147, 180, 150, 186, 640, 182, 85, 768, 167, 3~
## $ Region     <chr> "W", "NC", "S", "W", "NE", "NC", "S", "NC", "S", "NE", "NE"~
```

```
fit <- lm(Infections ~ Beds * Region, dataq5)
summary(fit)
```

```
##
## Call:
## lm(formula = Infections ~ Beds * Region, data = dataq5)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.2167 -1.2680 -0.1742  1.0825  9.3026
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.793438   0.868645  -0.913  0.36311
## Beds           0.038005   0.002514  15.119  < 2e-16 ***
## RegionNE      -1.226376   1.289732  -0.951  0.34385
## RegionS       -0.531393   1.176101  -0.452  0.65233
## RegionW        0.101267   1.347436   0.075  0.94023
## Beds:RegionNE  0.012096   0.003926   3.081  0.00263 **
## Beds:RegionS   0.001573   0.003571   0.441  0.66042
## Beds:RegionW  -0.004480   0.004643  -0.965  0.33679
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.885 on 105 degrees of freedom
## Multiple R-squared:  0.8912, Adjusted R-squared:  0.884
## F-statistic: 122.9 on 7 and 105 DF,  p-value: < 2.2e-16
```

The model: $Infections = \beta_0 + \beta_1 * Beds + \beta_2 * I(NE) + \beta_3 * I(S) + \beta_4 * I(W) + \beta_5 * Beds * I(NE) + \beta_6 * Beds * I(S) + \beta_7 * Beds * I(W)$

$= -0.793438 + 0.038005 * Beds - 1.226376 * I(NE) - 0.531393 * I(S) + 0.101267 * I(W) + 0.012096 * Beds * I(NE) + 0.001573 * Beds * I(S) - 0.004480 * Beds * I(W)$

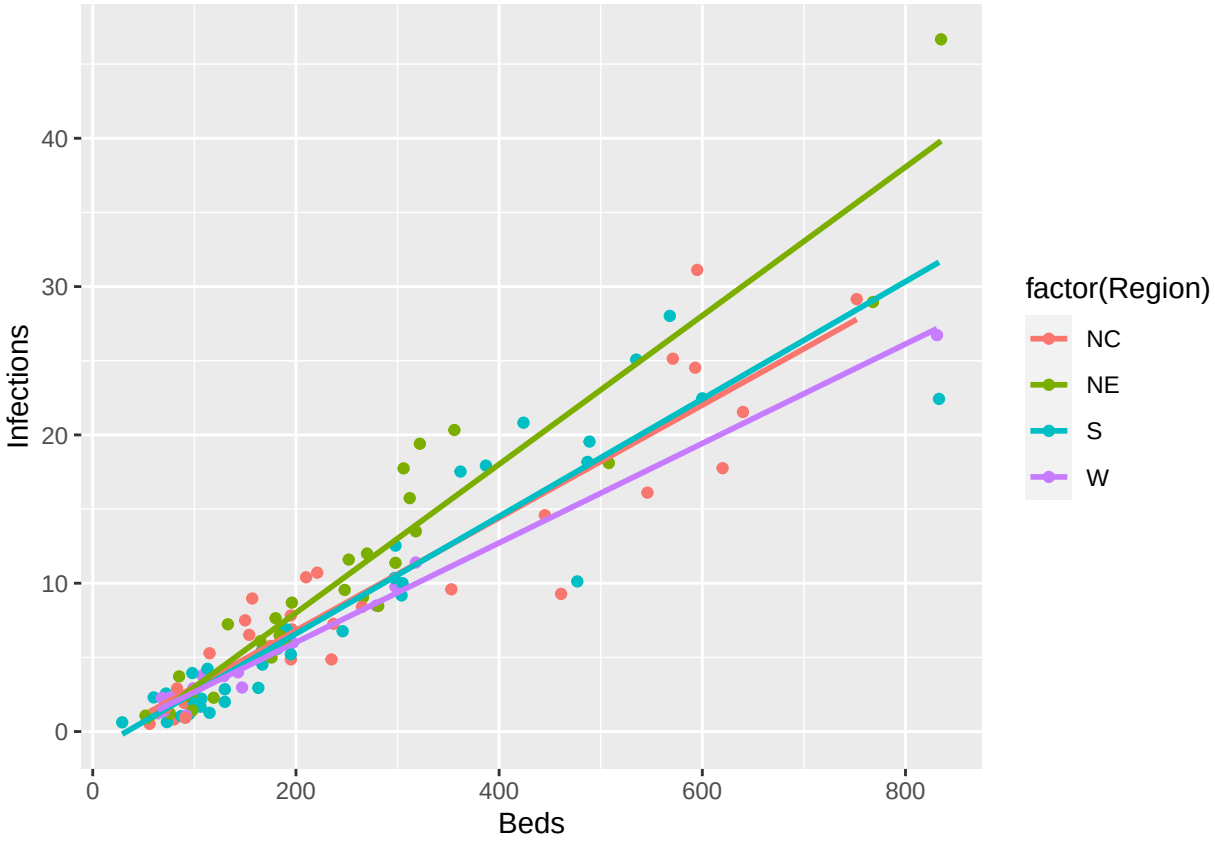
Estimate regression function for NE : $-2.019814 + 0.050101 * Beds$

Estimate regression function for S : $-1.324831 + 0.036432 * Beds$

Estimate regression function for W : $-0.692171 + 0.033525 * Beds$

Estimate regression function for NC : $-0.793438 + 0.038005 * Beds$

```
ggplot(dataq5, aes(x=Beds, y = Infections, color = factor(Region))) +
  geom_point() + geom_smooth(formula='y ~ x', method = 'lm', se = FALSE)
```



(b) $H_0 : \beta_5 = \beta_6 = \beta_7 = 0$

H_α : at least one of $\beta_5, \beta_6, \beta_7$ not equal to 0.

```
newfit <- lm(Infections ~ Beds + Region, dataq5)
anova(newfit,fit)
```

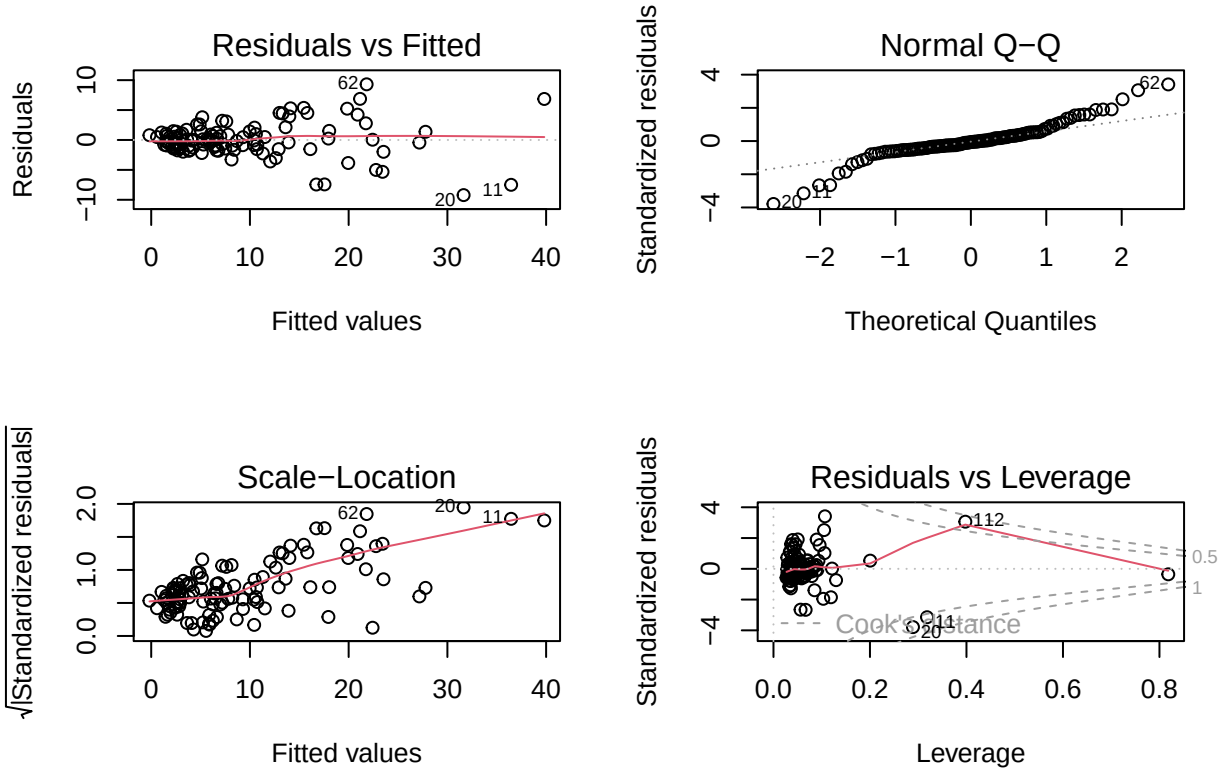
```
## Analysis of Variance Table
##
## Model 1: Infections ~ Beds + Region
## Model 2: Infections ~ Beds * Region
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    108 994.42
## 2    105 874.21  3    120.22 4.8131 0.003516 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For p-value is 0.003516, It is < 0.05, we reject H_0. That means at $\alpha = 0.05$, the slopes are not same at different regions.

(c) We will choose the model with different slope because this model is a full model with interaction.

(d)

```
par(mfrow = c(2,2))
plot(fit)
```



The standard assumptions is not satisfied. From the plot, we can see that the residuals is not normal (from the q-q plot, the residuals is not on normal line) and violates constant variance (The residuals on the left are denser).

It fits better. The constant variance are improved but The distribution is still not very uniform (The residuals on the left are denser). From the qq plot, we can see that the residuals more normal.